

[Open in app](#)

Parveen Khurana

124 Followers

[About](#)[Following](#)

Information Theory

 **Parveen Khurana** Jan 5, 2020 · 8 min read

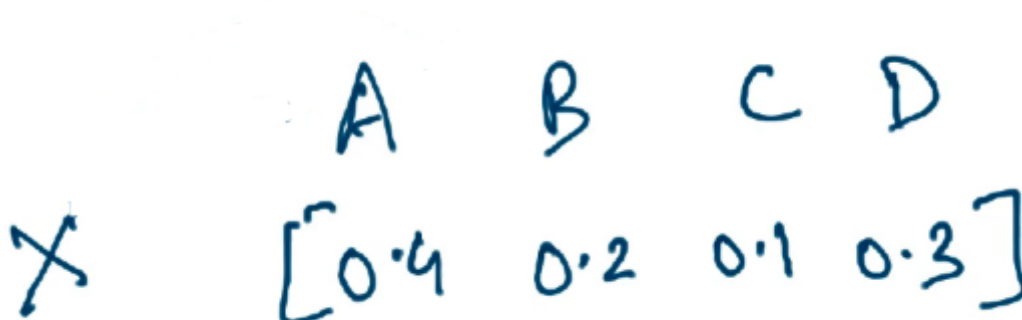
This article covers the content discussed in the Information Theory module of the [Deep Learning course](#) and all the images are taken from the same module.

In the [previous article](#), we discussed how in the context of classification we can represent true output and the predicted output as a probability distribution. In this article, we will see a new way to compute the difference between the two distributions or to say compute the loss value for the distributions.

Expectation:

Once again, think of a situation where we have 4 teams A, B, C and D and there is random variable X which denotes the winning team.

And say based on some past experience, we come up with the probability distribution of these teams winning the tournament as below:



Handwritten probability distribution for teams A, B, C, and D:

	A	B	C	D
X	0.4	0.2	0.1	0.3

[Open in app](#)

wins you lose some 8k, if team D wins, you get 5k):

A	B	C	D
0.4	0.2	0.1	0.3
10K	2K	-8K	5K

So, there is some probability associated with each of the random variables and some profit/loss(Gain) associated with each of the possible values of the random variables. The expected return we can compute as a summation of the product of the probability and the associated gain for all the values that the random variable can take:

$$E_x[p] = \sum_{i \in \{A, B, C, D\}} p(x=i) G(x=i)$$

$$= 0.4(10K) + 0.2(2K) + 0.1(-8K) + 0.3(5K)$$

[Open in app](#)

So, the above value is the Expected Return value as per the probability distribution and the Gain/Loss associated with each of the values that the random variable can take.

Information Content

The intuition behind Information Content:

Let's take an event which is where does the sun rises so let's say the random variable is where does the sunrise? And it can take 4 values: East, West, North, South.

And now if we tell that the Sun rose in the East today, then there is nothing much information gain because this is a certain/sure event that the sun rises in the East which happens with probability 1. So, there is no information that we are gaining from this.

Let's take another event. Suppose we tell you today there is going to be a Storm, so let's say the random variable Y can take on the value of Storm or No Storm.

Now if we get to know that today there is going to be a storm, in this case, the information gain would be very high because regularly there is no storm. A storm is something like a very very rare event and if we tell you about the rare event, then the information gain is really high. There is a lot of surprise for you and the surprise leads to information gain.

Based on the above, we could say Information Gain is directly proportional to the Surprise in the event.

$$IC(x \in S) \propto \text{surprise}$$

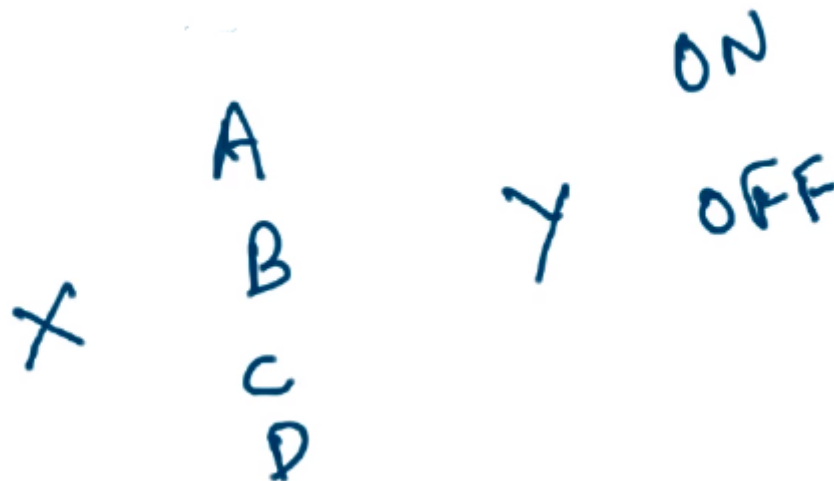
Now since we are discussing probability, so the surprise equivalent in probability can be deduced as: A surprising event is the one which has a low probability. So, low

[Open in app](#)


$$I(X=S) \propto \text{surprise} \propto \frac{1}{P(X=S)}$$

The more surprising the event, the less the probability and the less the probability the more information we gain by knowing about it.

Now let's consider another scenario in which a tournament is going on and the random variable X tells which team is going to win from A, B, C, or D and there is this another random variable Y which tells whether the A.C in this room is ON or OFF.



Now suppose we tell you that team B won and the A.C in this room is ON. The first thing to note here is that these **two random variables are independent**, A.C being on or off in this room has no bearing on the match outcome and similarly who is going to win the match has no bearing on whether the A.C in this room is on or off. So, these are completely independent events. So, if we tell about two independent contents, then what is Information Content going to be:

[Open in app](#)


content by knowing both of these (independent events) should just be the summation of the individual information content (we know that the information gain is a function of the probability of the event and is inversely proportional to probability)

$$IC(P(X \cap Y)) = IC(P(X)) + IC(P(Y))$$

So, now we have this interesting situation, we have this Information Content as a function such that it satisfies the below criteria:

$$f(a \cdot b) = f(a) + f(b)$$

The above criteria is satisfied by the **Logarithmic family of functions**. So, in essence, we have:

$$IC(X=A) = \log\left(\frac{1}{P(X=A)}\right)$$

which we can re-write as:

Open in app



$$\begin{aligned}
 I(x=A) &= -\log(P(x=A)) \\
 &= \log 1 - \log(P(x=A)) \\
 I(x=A) &= -\log_2 P(x=A)
 \end{aligned}$$

Entropy:

Let's say we have a random variable X which can take on values A, B, C, D .

X	$P(X=?)$	$I(X=?)$
A	$P(x=A)$	$-\log_2 P(x=A)$
B	$P(x=B)$	$-\log_2 P(x=B)$
C	$P(x=C)$	$-\log_2 P(x=C)$
D	$P(x=D)$	$-\log_2 P(x=D)$

Now based on the above table, we can say that the Expected Gain is given by:

$$= \sum P(x=i) \text{Gain}(x=i)$$

[Open in app](#)


Entropy($H(x)$) is the expected information content of a random variable and is given as:

$$H(x) = - \sum_{i \in A, B, C, D} p(x=i) \log_2 p(x=i)$$

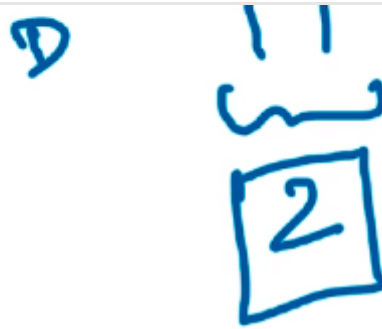
Relation to Number of Bits required to transmit a message

Suppose there is this $X(\text{message})$ that we would like to transmit and the message can take on 4 values A, B, C, or D.



We can use 2 bits to transmit 4 messages:

A	0 0
B	0 1

[Open in app](#)


So, in all, for every message we are transmitting, we are using 2 bits.

In this case, we are assuming that all of these messages are equally likely that is their probability is $1/4$. So, this is the distribution that we are assuming for Random Variable taking on any of the 4 values it can take.

$P(X=?)$

A	0 0	$\frac{1}{4}$
B	0 1	$\frac{1}{4}$
C	1 0	$\frac{1}{4}$
D	1 1	$\frac{1}{4}$

Let's see the information content for each of the messages:

[Open in app](#)


$P(X=?)$ $-\log_2 P(X=?)$
 $-\log_2 \frac{1}{4} = 2$

A	0 0	$\frac{1}{4}$	2
B	0 1	$\frac{1}{4}$	2
C	1 0	$\frac{1}{4}$	2
D	1 1	$\frac{1}{4}$	2

$\underbrace{\hspace{1.5cm}}$
2

Now we can make this connection that the **number of bits required to transmit a message is equal to the Information Content of that message.**

$P(X=?)$ $-\log_2 P(X=?)$
 $-\log_2 \frac{1}{4} = 2$

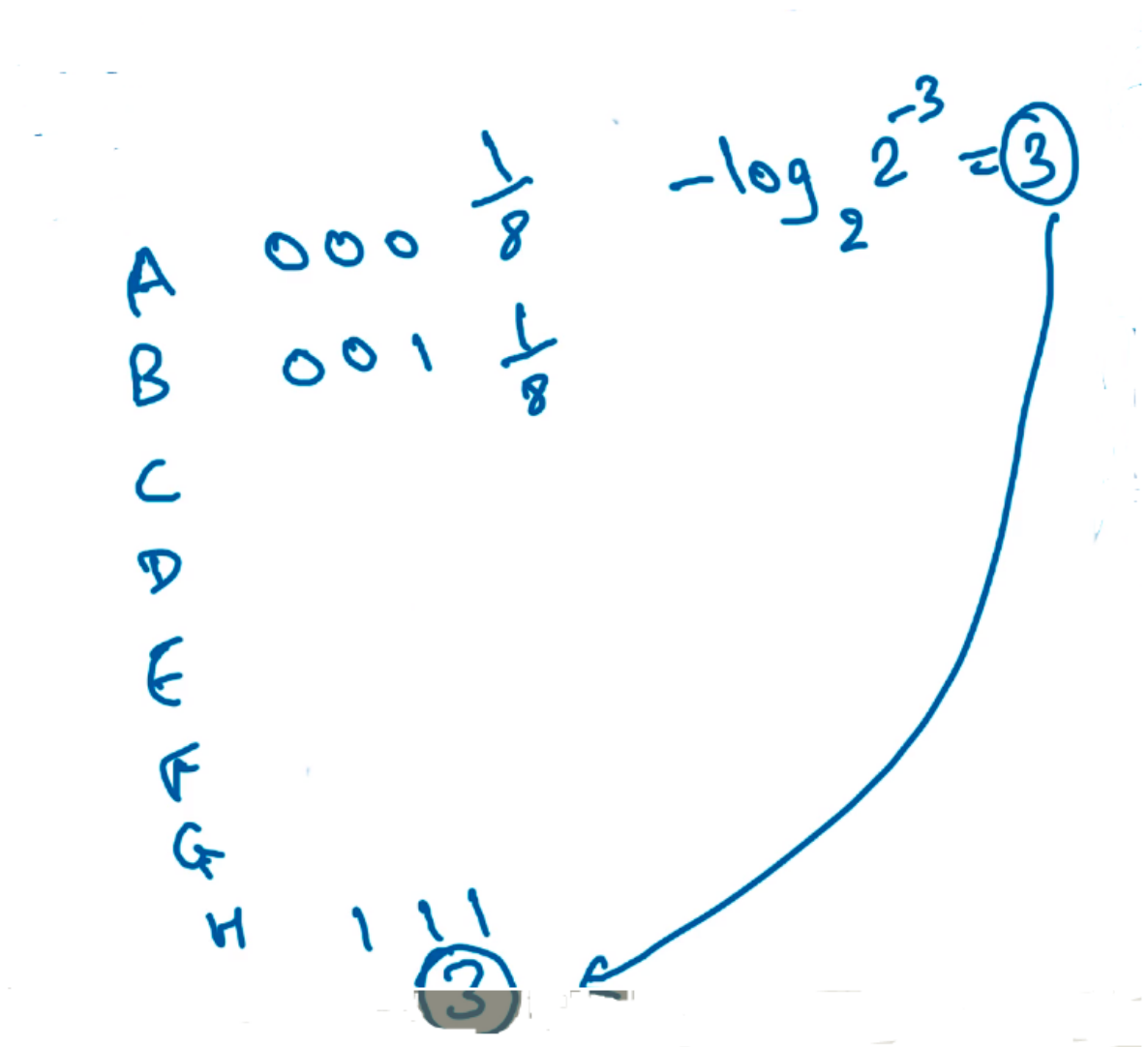
A	0 0	$\frac{1}{4}$	2
B	0 1	$\frac{1}{4}$	2
C	1 0	$\frac{1}{4}$	2

\curvearrowright (A curved arrow points from the circled '2' in the calculation to the '2' in the rightmost column of the table.)

[Open in app](#)


2

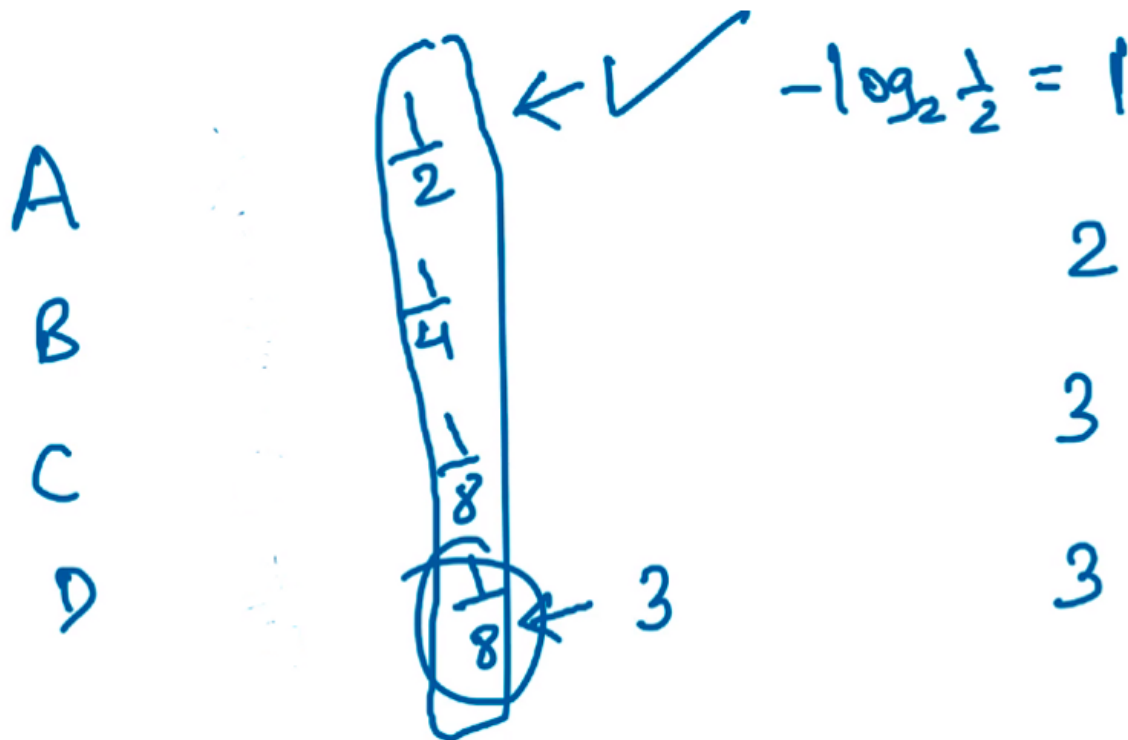
Let's consider this for 8 messages(A, B, C, D, ..., H)



Now if we want to send a continuous stream of messages and we want to minimize the number of bits required to do so. Let's say we have a different distribution, say message A is the most frequent message and so on.

[Open in app](#)


of bits required to send each of the messages would be (assuming the probability/frequency of each message as in the below image):



The only way we can say this strategy is better is if on average we end up using less no. of bits (even if we are using 3 bits for less frequent messages)

$$\begin{array}{cccc}
 A & B & C & D \\
 \left[\frac{1}{2} & \frac{1}{4} & \frac{1}{8} & \frac{1}{8} \right] \\
 \\
 I_c & \left[1 & 2 & 3 & 3 \right]
 \end{array}$$

[Open in app](#)


$$H(X) = - \sum p_i \log_2 p_i$$

$$\frac{1}{2} (1) + \frac{1}{4} (2) + \frac{1}{8} (3) + \frac{1}{8} (3)$$

$$= \underline{\underline{1.75}}$$

So, the entropy of a random variable tells us the average number of bits required to send that random variable.

KL Divergence and Cross-Entropy

Let's say X is the random variable and y is the true distribution of the random variable

	A	B	C	D	
X	y_1	y_2	y_3	y_4	y
$I_C(X)$	$[-\log y_1]$	$[-\log y_2]$	$[-\log y_3]$	$[-\log y_4]$	

[Open in app](#)


associated with these messages given by y and we don't know this true distribution in advance and you are looking at some of these messages and you are estimating a y_{hat} (predicted distribution), so let's say the predicted distribution looks like:

$$\begin{array}{cccc}
 & A & B & C & D \\
 \times & [y_1 & y_2 & y_3 & y_4] & \textcircled{y} \\
 IC(x) & [-\log y_1 & -\log y_2 & -\log y_3 & -\log y_4] \\
 & [\hat{y}_1 & \hat{y}_2 & \hat{y}_3 & \hat{y}_4] & \hat{y}
 \end{array}$$

The true entropy of the random variable would be:

$$-\sum y_i \log y_i$$

But we have predicted some distribution/probability for the random variable and as per that the information content would look like:

$$\begin{array}{cccc}
 & A & B & C & D \\
 \times & [y_1 & y_2 & y_3 & y_4] & \textcircled{y} \\
 IC(x) & [-\log y_1 & -\log y_2 & -\log y_3 & -\log y_4]
 \end{array}$$

[Open in app](#)


$$\hat{I}_C(x) [\underline{-\log \hat{y}_1}, \underline{-\log \hat{y}_2}, \dots]$$

The actual messages at the destination would come as per the true distribution y , so the predicted no. of bits that we end up using is going to be:

$$-\sum y_i \log \hat{y}_i$$

the value associated with each of the values that the random variable can take is

$$\log \hat{y}_i$$

because that is what we have estimated but the messages actually are going to come by probability as per true distribution.

So, if we knew the true distribution, then the number of bits that we would require would have been

$$-\sum y_i \log y_i$$

This we can call as Entropy.

[Open in app](#)


$$-\sum y_i \log \hat{y}_i$$

The above is the cross-entropy between two distribution y and y_{hat} .

$$-\sum y_i \log \hat{y}_i$$

\uparrow
 $H_{y, \hat{y}}$

So, now we have the bits required to transmit a message as per the true distribution and as per the estimated distribution and we can now compute the difference (this difference is known as KL Divergence) between these two as:

$$KLD(y \parallel \hat{y}) = -\sum y_i \log \hat{y}_i + \sum y_i \log y_i$$

And this KL Divergence provides a way of computing the difference between two distributions.

Open in app



About Write Help Legal

Get the Medium app

